

# The SPECIALIST NLP Tools

Dr. Chris J. Lu

The Lexical Systems Group

NLM. LHNCBC. CGSB

June, 2010

# Table of Contents

- Introduction
- Lexical Tools
  - Lvg
  - Norm
- Text Categorization Tools
- Questions

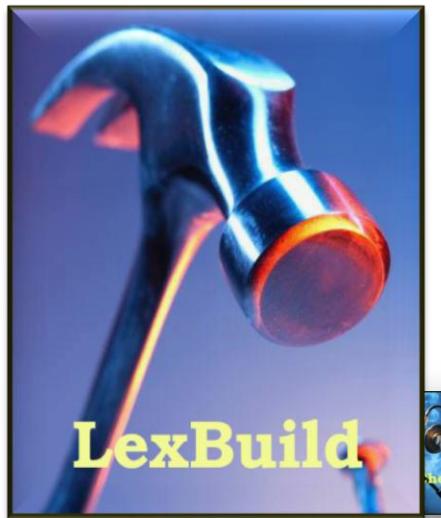
# Introduction



# Introduction



# Introduction - LB



check



LexAccess



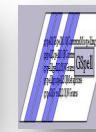
Ten  
Numbers



SCRT



Lexical Tool



Text Tools



dragger



VTT



TC  
JDI  
WSD

# Introduction - Lexicon



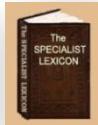
# Introduction - LC



# Introduction - LA



# Introduction - Numbers



# Introduction - SCRT



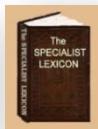
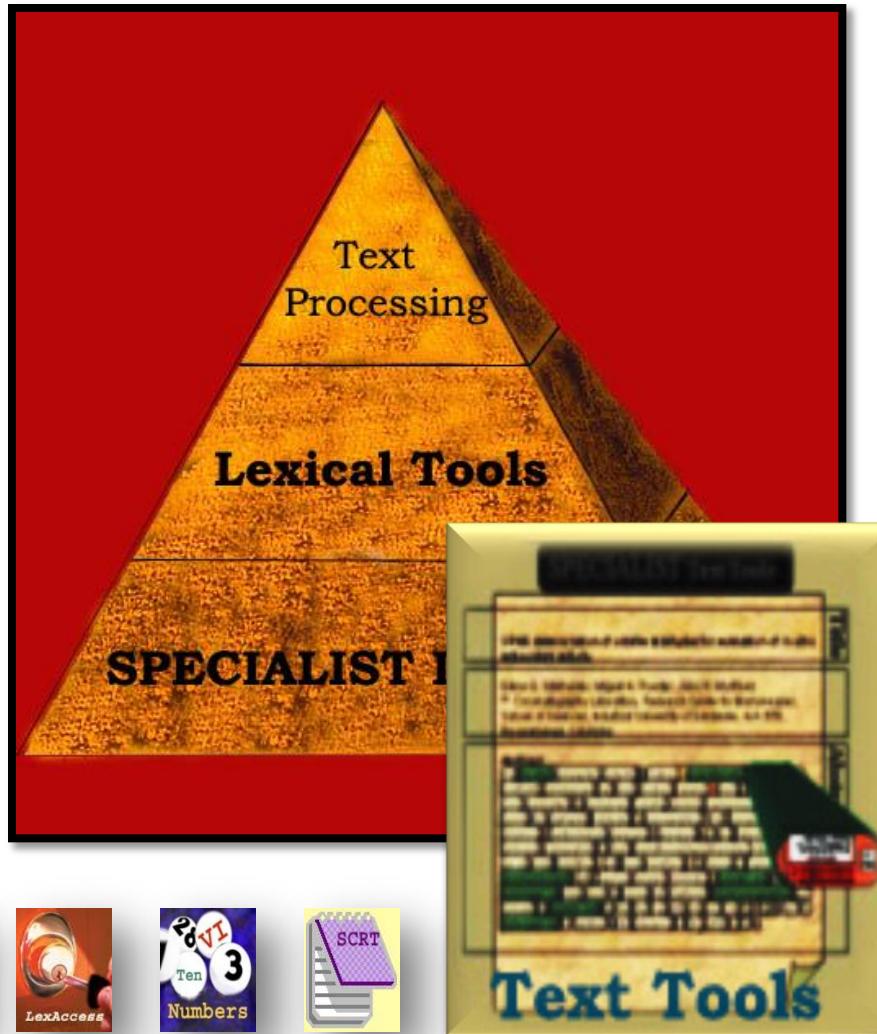
# Introduction – Lexical Tools



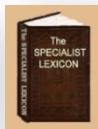
Lexical Tool



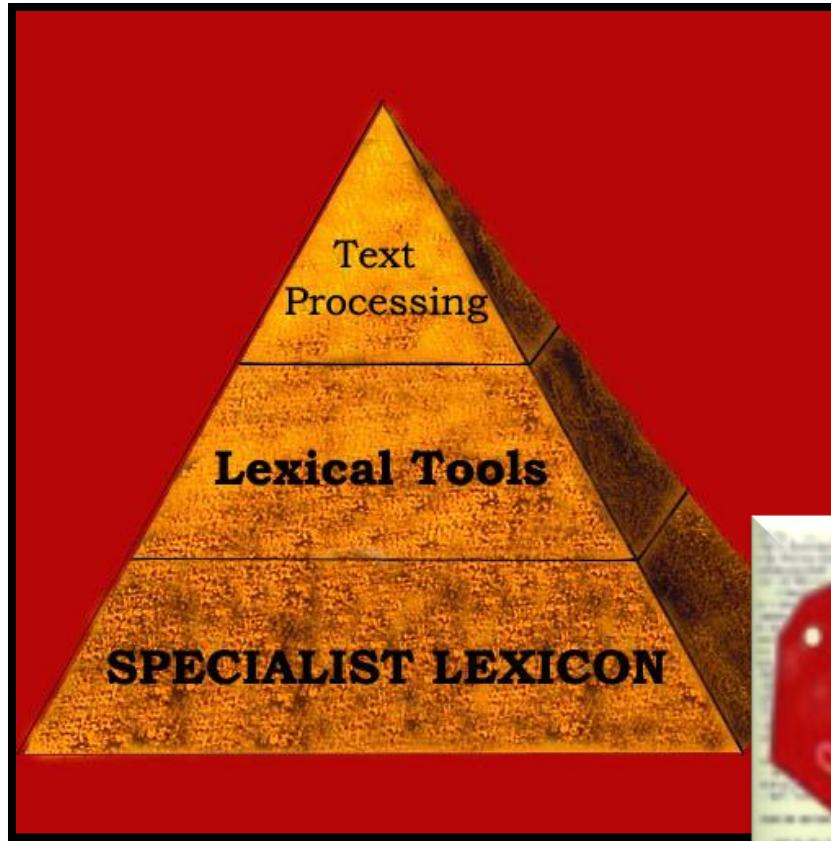
# Introduction – Text Tools



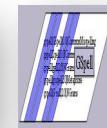
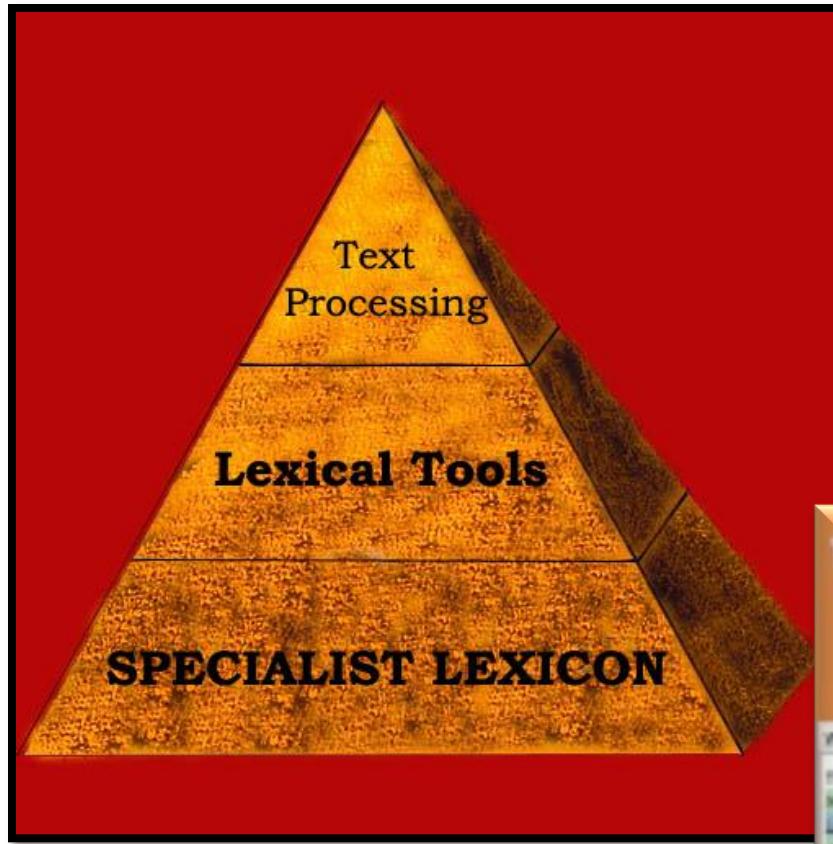
# Introduction - GSpell



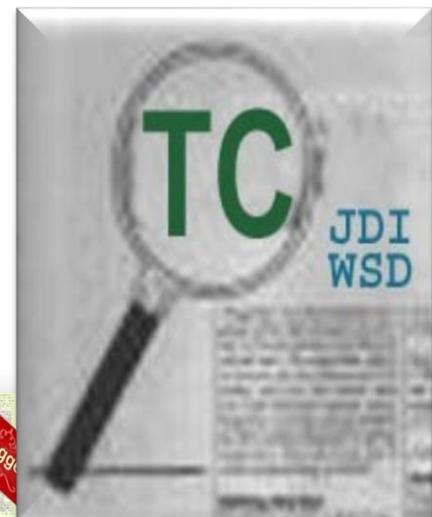
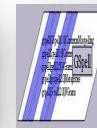
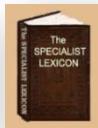
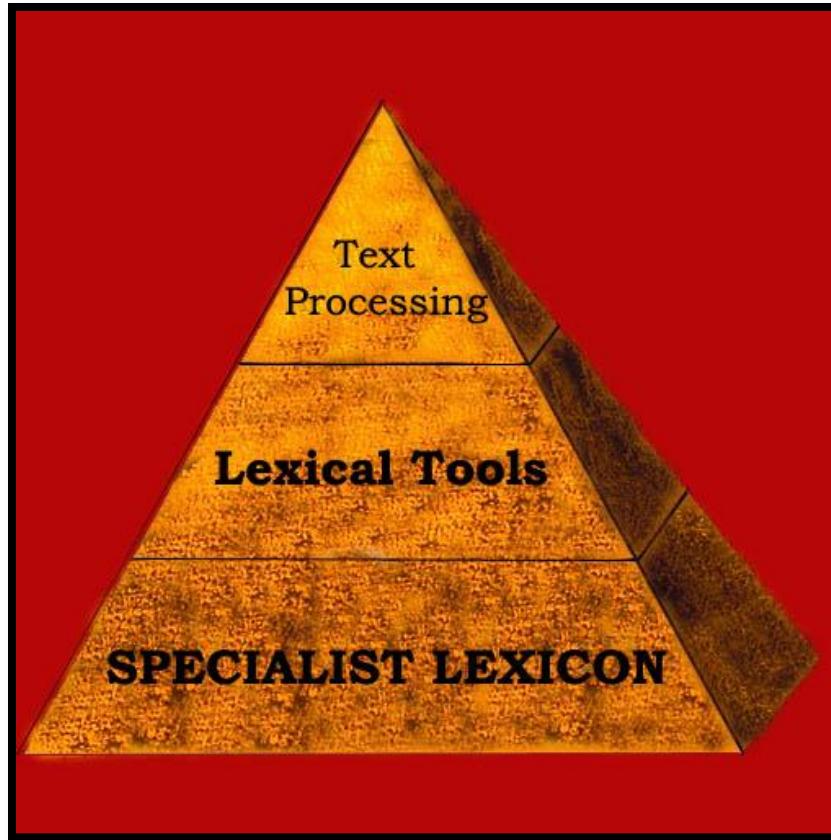
# Introduction - dTagger



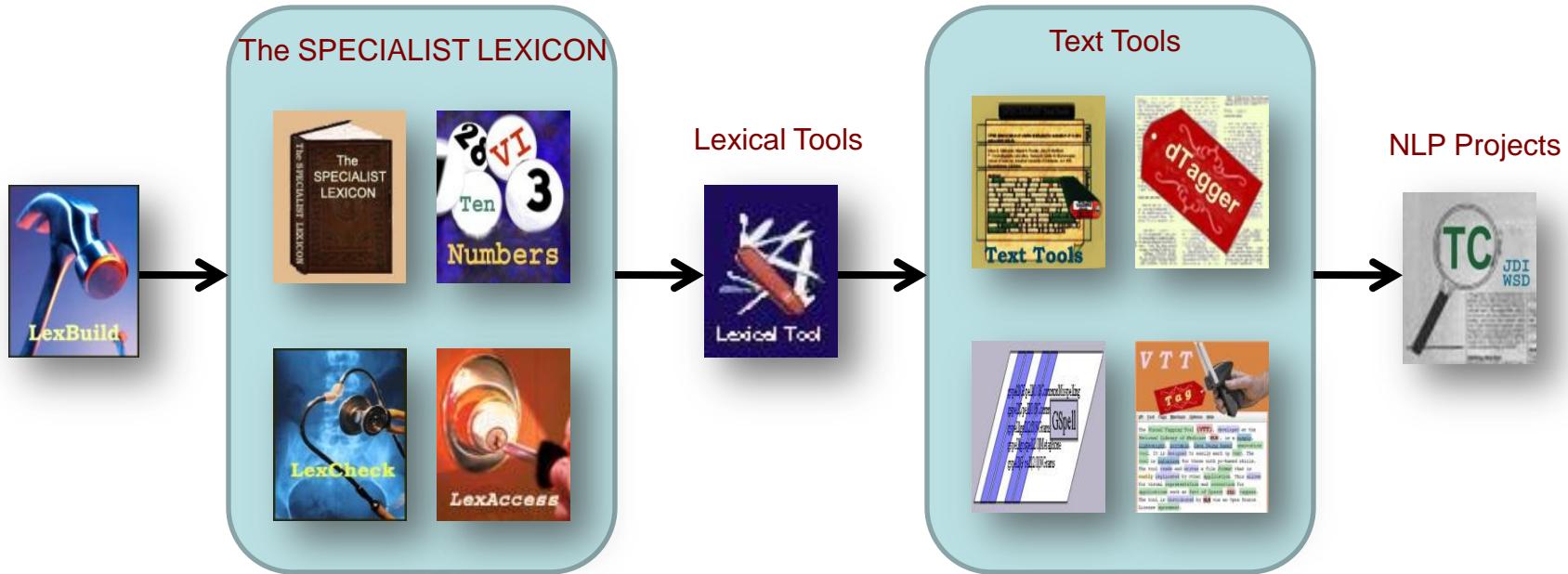
# Introduction - VTT



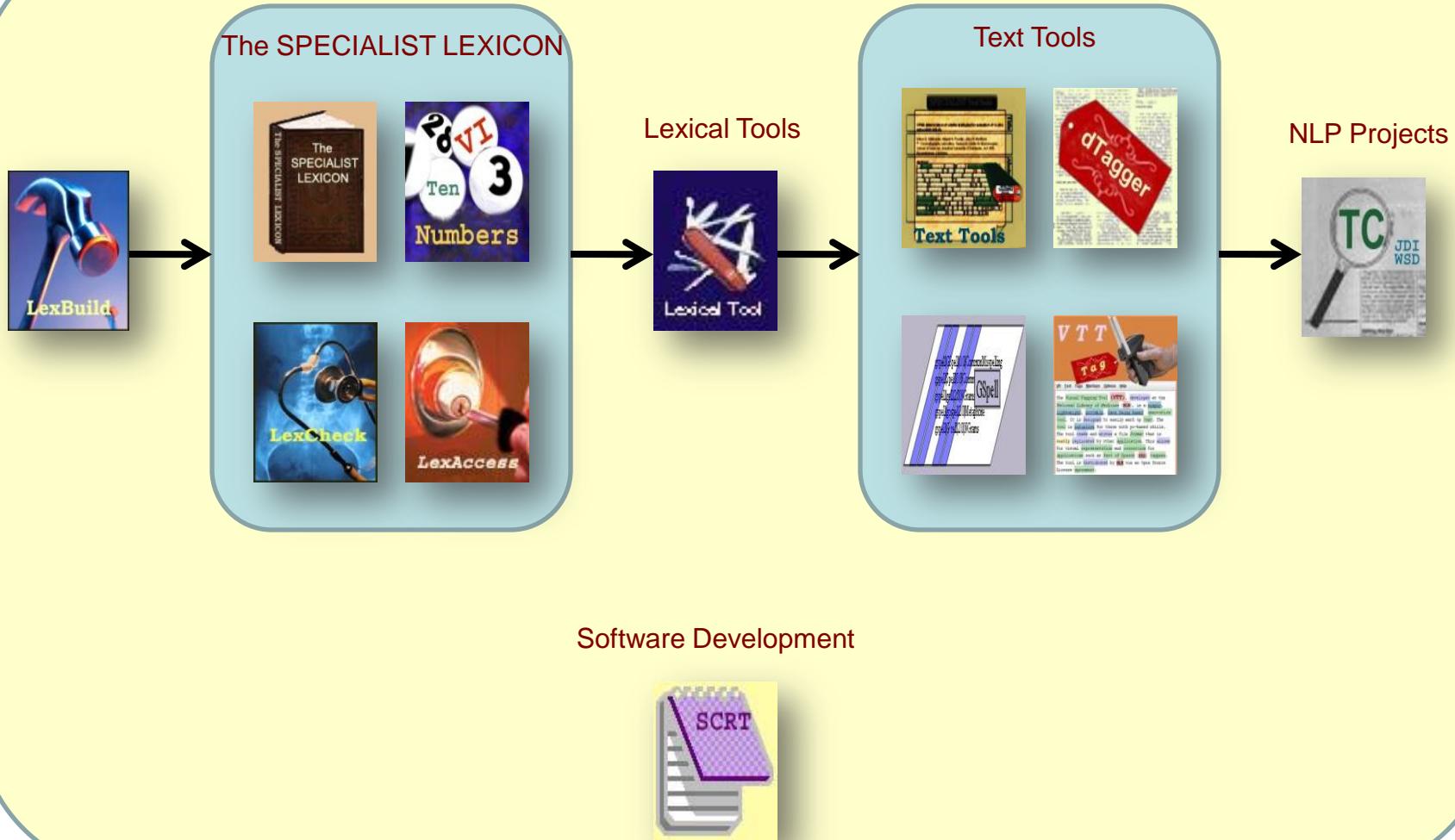
# Introduction - TC



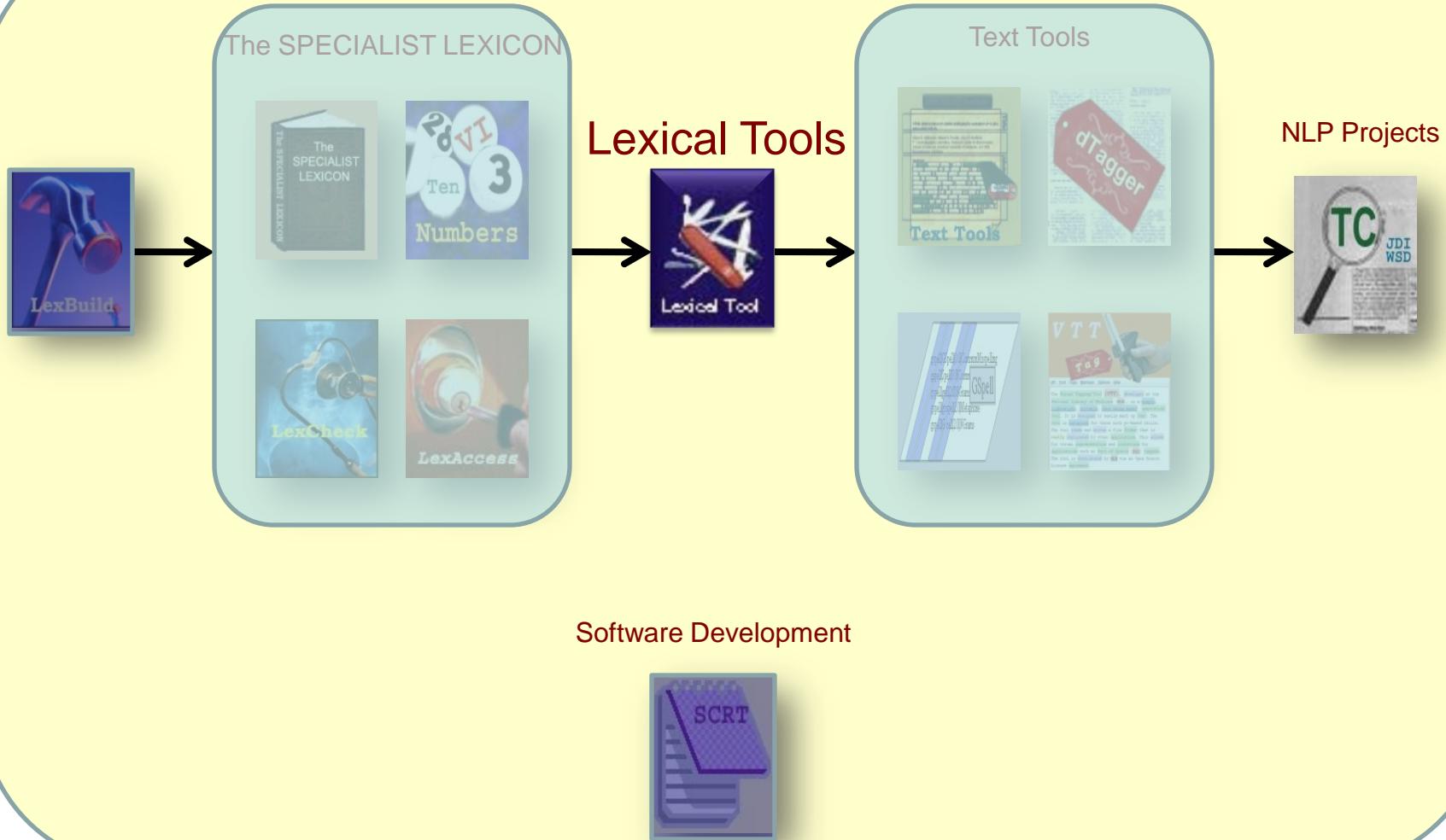
# Introduction – NLP Tools



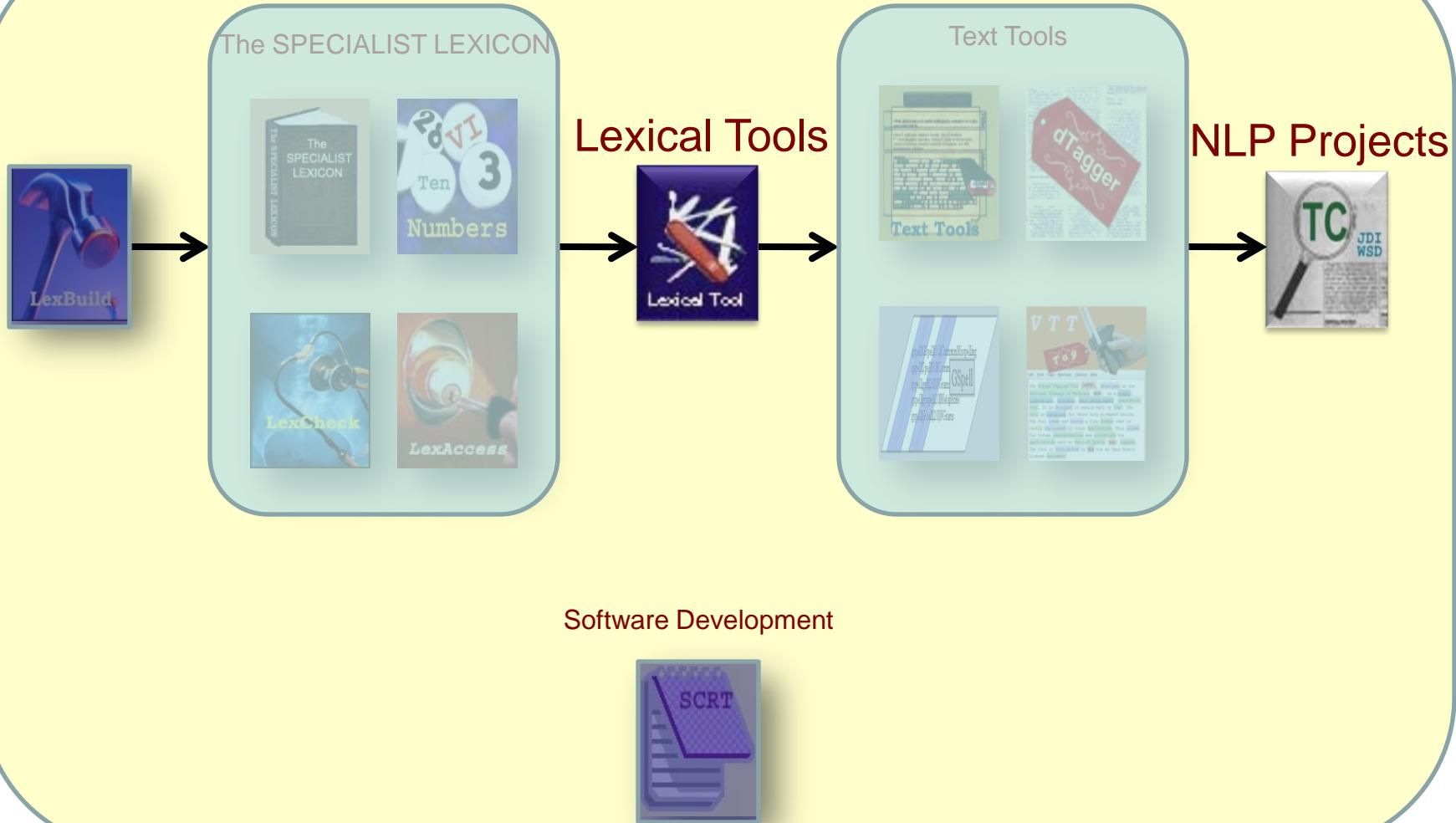
# Introduction – NLP Tools



# Introduction – NLP Tools



# Introduction – NLP Tools



# Lexical Tools

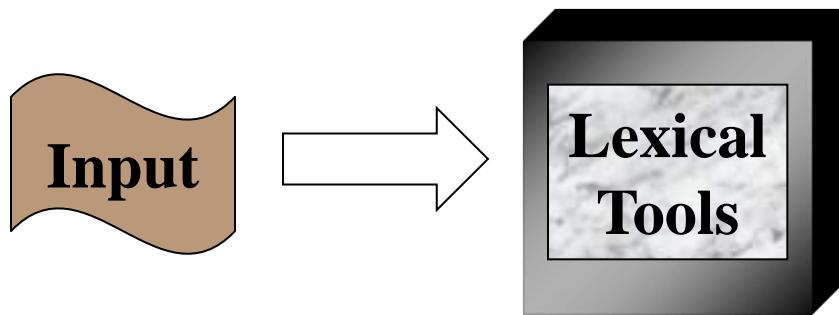


# Lexical Tools



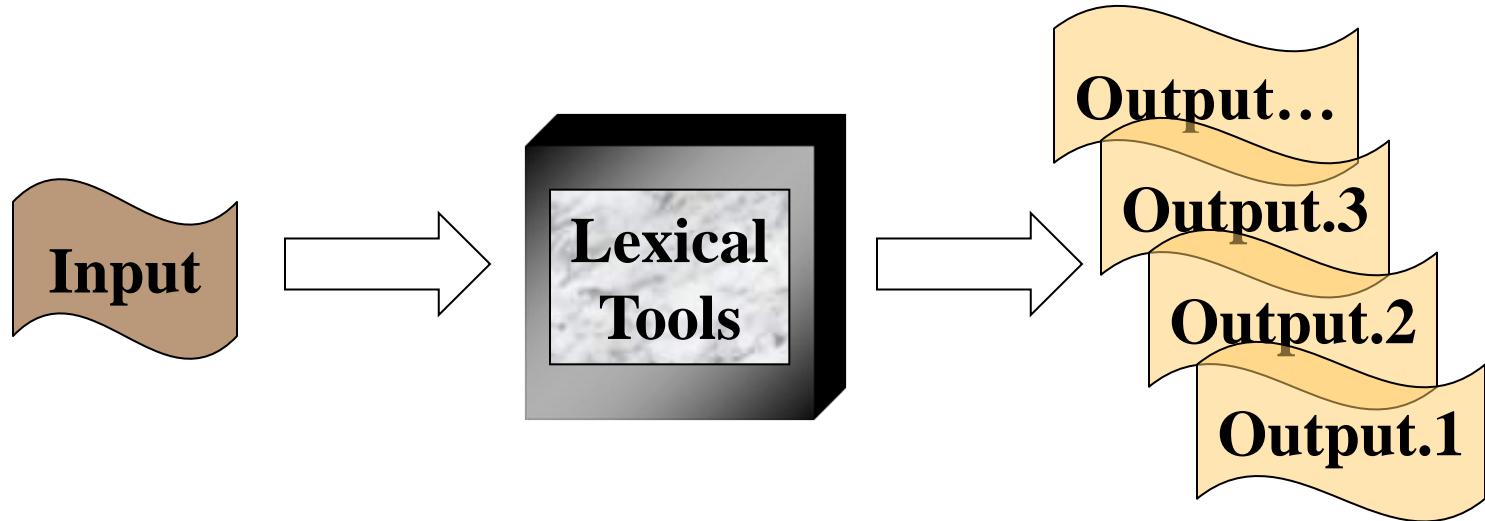
- A suite of text utilities

# Lexical Tools



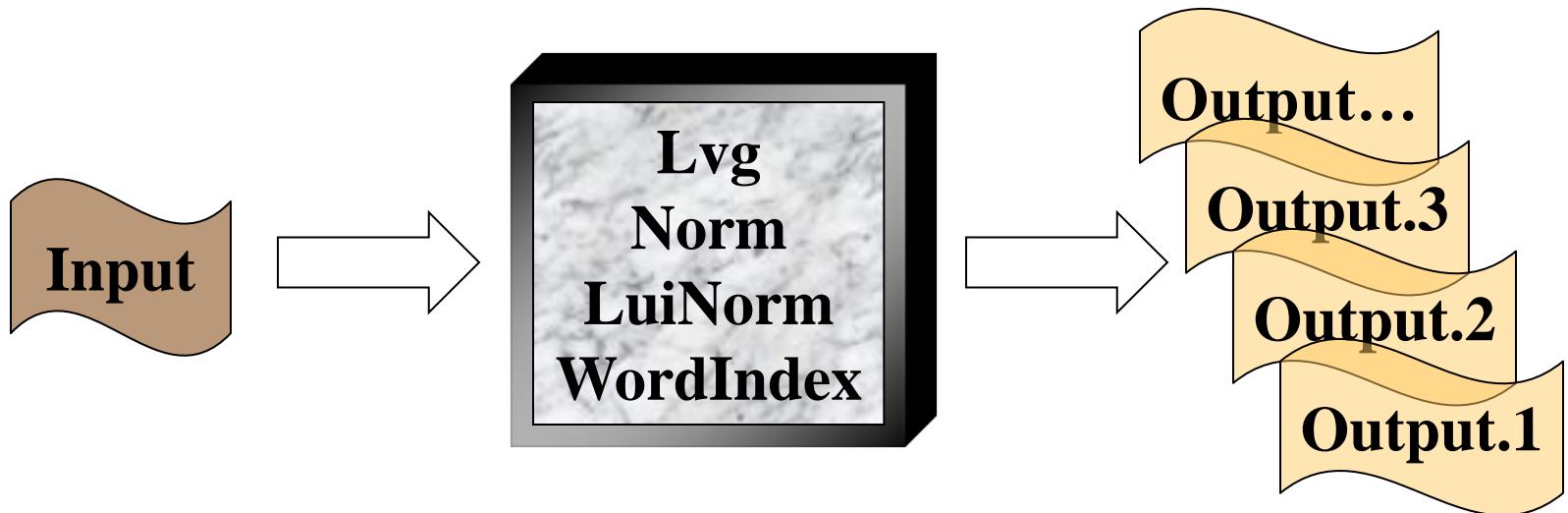
- A suite of text utilities take the given input

# Lexical Tools



- A suite of text utilities that generate, mutate, and filter out lexical variants from the given input

# Four Tools



# Tool Types

- Command line tools
  - lvg (Lexical Variants Generation)
  - norm
  - luiNorm
  - wordInd
- Lexical Gui Tool (Igt)
- Web Tools
- Java API's

# Functions

- Used in nature language processing for
  - aggressive text pattern matching
  - creating normalized and expanded terms
  - making word, term, phrase indexes
  - matching queries with indexed entries
  - increasing recall and/or precision

# Facts

- Release annually
- Free distributed with open source code
- 100% Java (since 2002)
- Run on different platforms
- One complete package
- Documents & supports

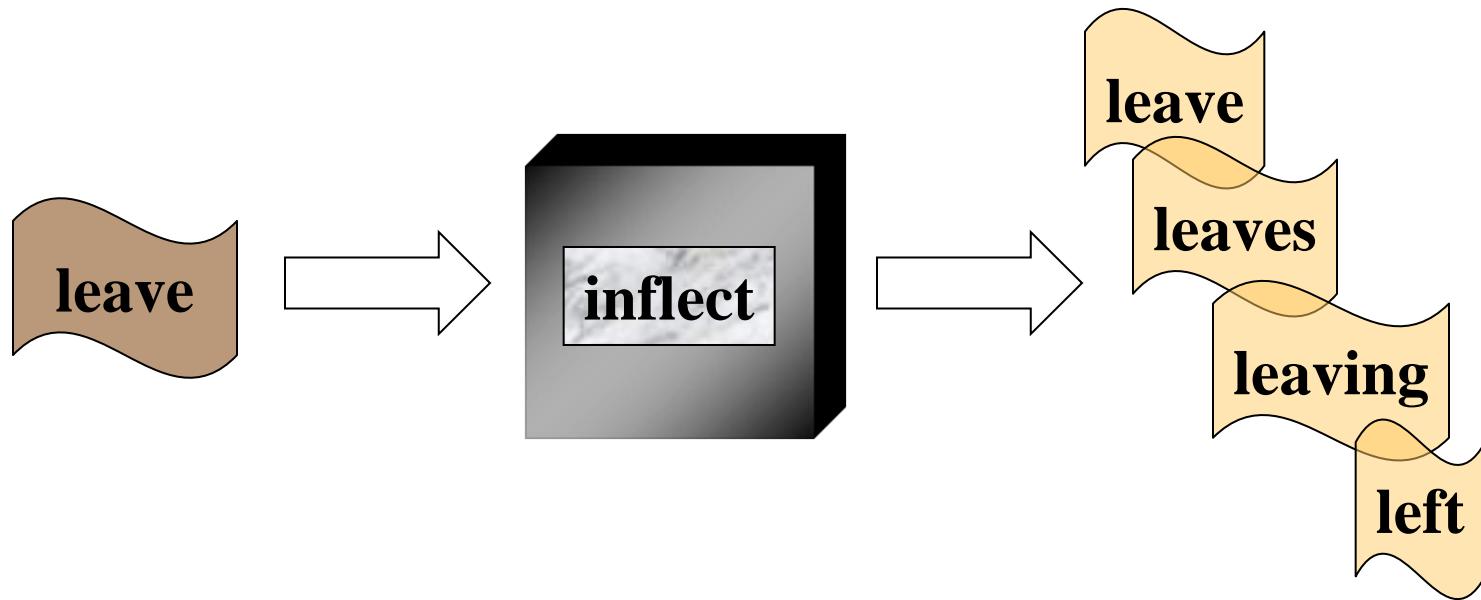
# Lexical Variants Generation



# **LVG - 2010**

- 62 flow components
- 37 options
  - input filter options (3)
  - global behavior options (13)
  - flow specific options (2)
  - output filter options (19)

# Flow Components

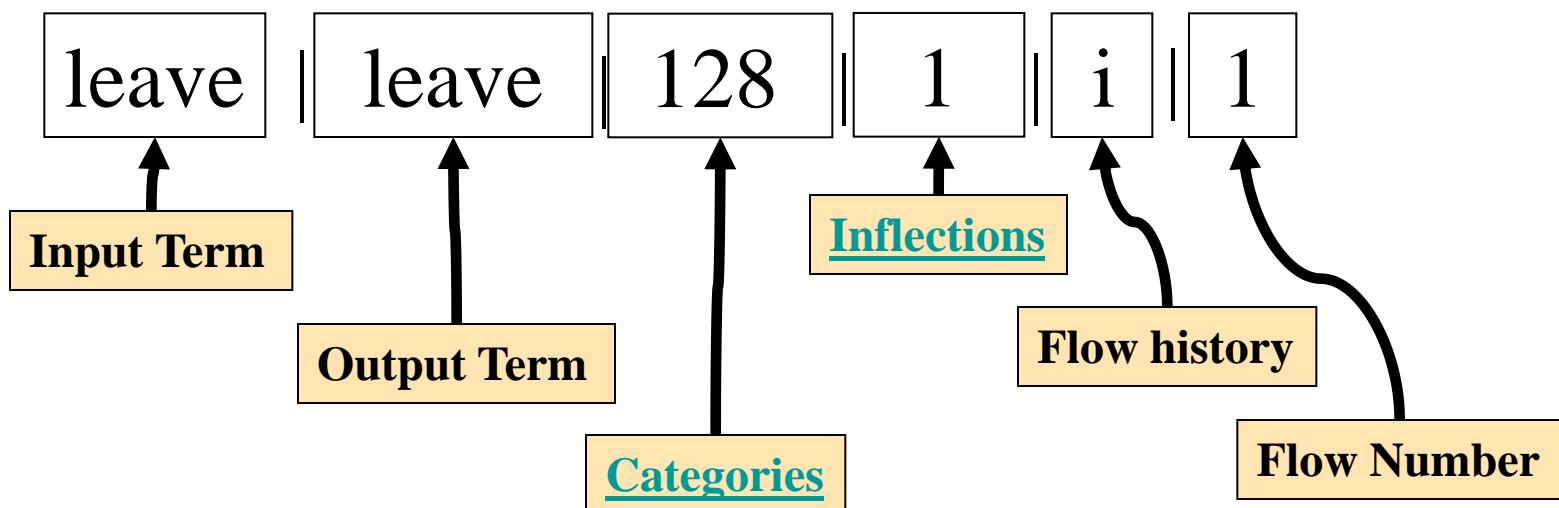


# Command Line Tool

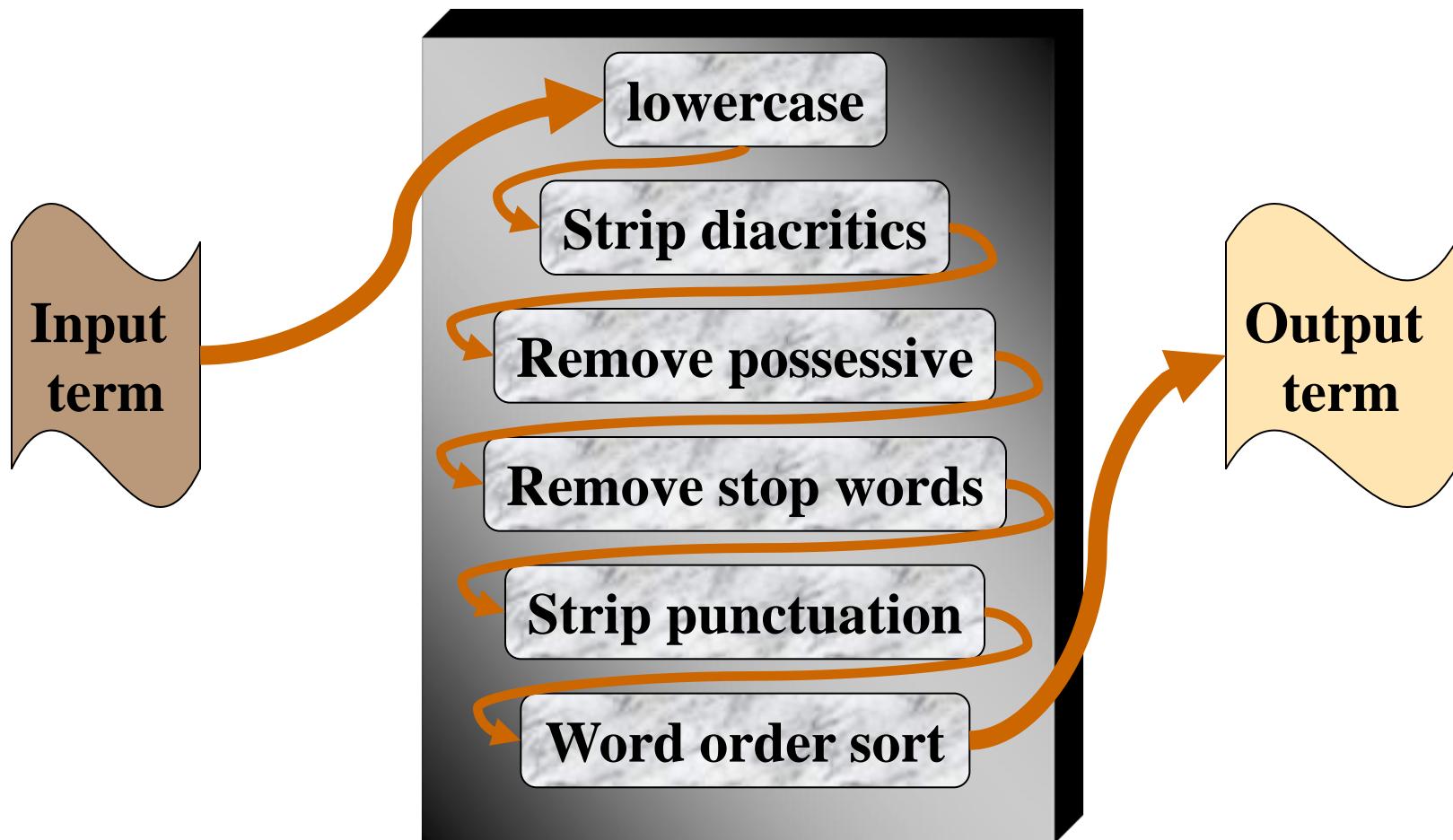
```
> lvg -f:i  
leave  
leave|leave| | |i|1|  
leave|leave|128|512|i|1|  
leave|leaves|128|8|i|1|  
leave|left|1024|64|i|1|  
leave|left|1024|32|i|1|  
leave|leave|1024|1|i|1|  
leave|leave|1024|262144|i|1|  
leave|leave|1024|1024|i|1|  
leave|leaves|1024|128|i|1|  
leave|leaving|1024|16|i|1|
```

# Fielded Output

> lvg –f:i  
leave



# A Serial Flow

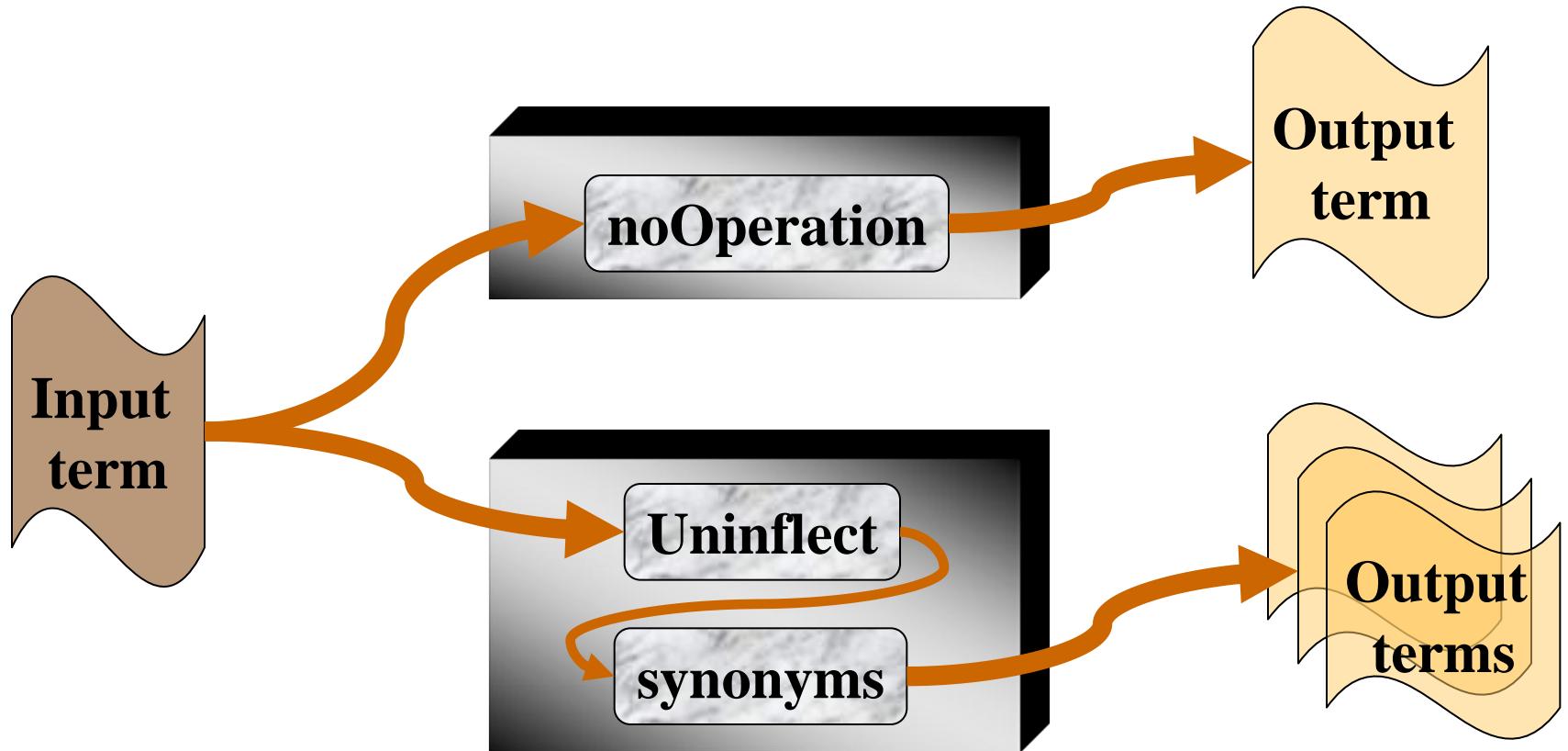


- Flow components can be arranged so that the output of one is the input to another.

# A Serial Flow - Example

```
> lvg -f:l:q:g:t:p:w  
The Gougerot-Sjögren's Syndrome  
The Gougerot-Sjögren's Syndrome |  
gougerotsjogren syndrome | 2047 |  
16777215 | 1+q+g+t+p+w | 1 |
```

# Parallel Flows

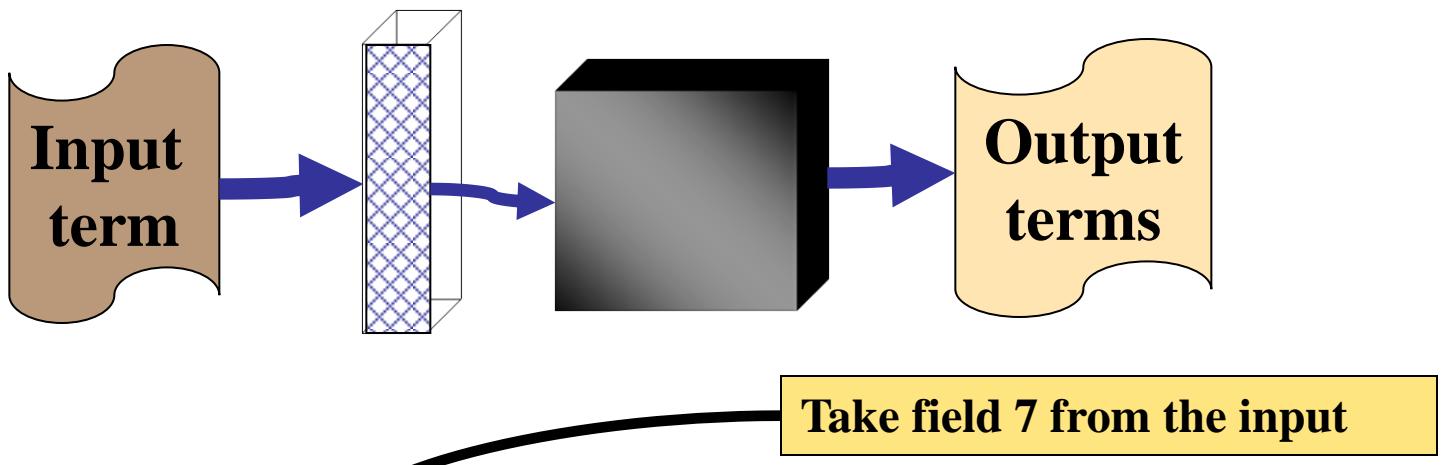


- Multiple flows can be defined

# Parallel Flows - Example

```
> lvg -f:n -f:B:y  
ear  
ear|ear|2047|1048575|n|1|  
  
ear|aural|1|1|B+y|2|  
ear|auricularis|1|1|B+y|2|  
ear|otic|1|1|B+y|2|  
ear|otor|1|1|B+y|2|
```

# Input Filter Options



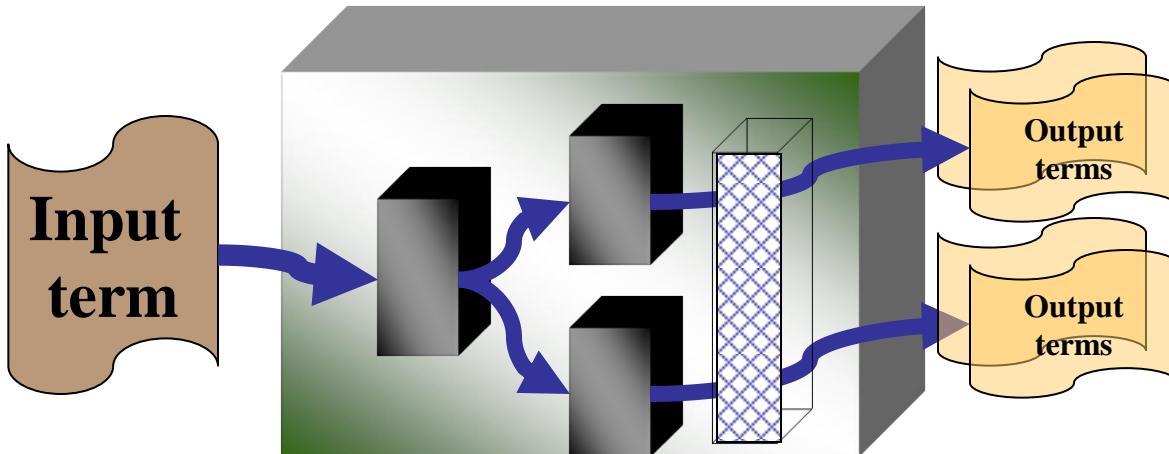
```
> lvg -f:u -t:7 -F:8:6
```

C0035440 | ENG | S | L0035434 | VW | S0003894 |

Rheumatic carditis, acute

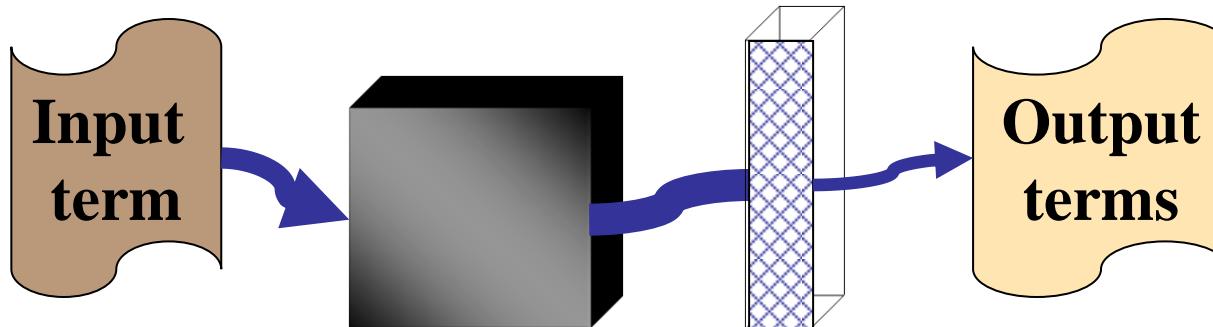
*acute Rheumatic carditis|S0003894*

# Global Behavior Options



```
> lvg -f:L -f:E -s:"\\" Change separator to "\"
otitis
otitis\otitis\128\513\L\1
otitis\E0044452\128\513\E\2
```

# Output Filter Options



> lvg -f:L **-SC -SI**

hot

hot|hot|<adj+verb>|<base+positive+infinitive+pres1p23p>|L|1|

→ **-SC -SI**

Show the category and inflection names

# Norm

- Composed of 11 Lvgl flow components to abstract away from:
  - case
  - punctuation
  - possessive forms
  - inflections
  - spelling variants
  - stop words
  - Diacritics, ligatures & symbols (Unicode to ASCII)
  - word order

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

# Norm

Hodgkin's Diseases, NOS

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

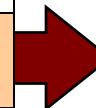
B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order



Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

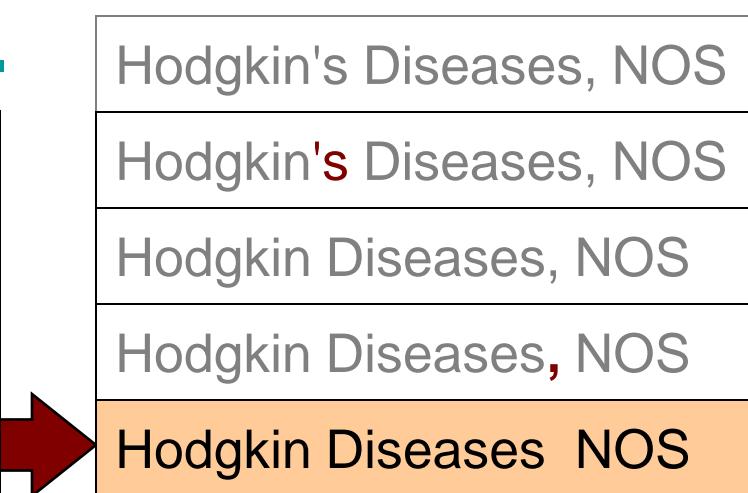
B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order



# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

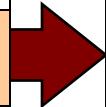
Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases



# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases

hodgkin diseases

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases

hodgkin diseases

hodgkin disease

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

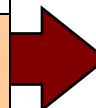
Hodgkin Diseases NOS

Hodgkin Diseases

hodgkin diseases

hodgkin disease

hodgkin disease



# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

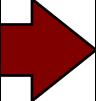
Hodgkin Diseases

hodgkin diseases

hodgkin disease

hodgkin disease

hodgkin disease



# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases

hodgkin diseases

hodgkin disease

hodgkin disease

hodgkin disease

hodgkin disease

# Norm

q0: map Unicode symbols to ASCII

g: remove genitives

rs: remove parenthetic plural forms

o: replace punctuation with spaces

t: strip stop words

l: lowercase

B: uninflect each words in a term

Ct: retrieve citations

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order

Hodgkin's Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases

hodgkin diseases

hodgkin disease

hodgkin disease

hodgkin disease

hodgkin disease

disease hodgkin

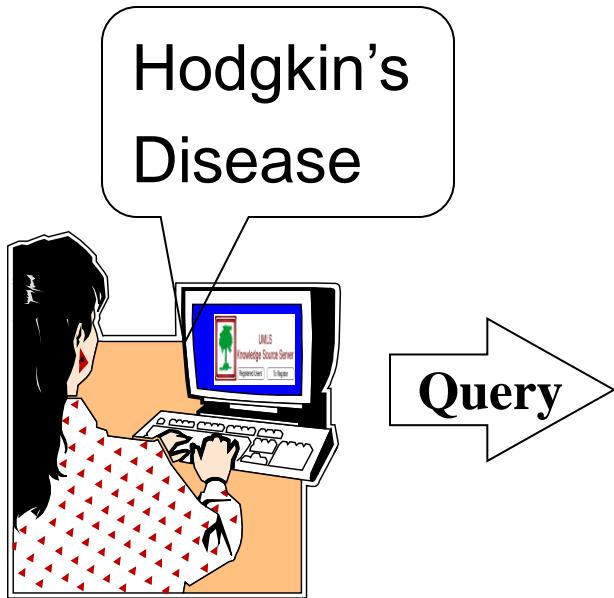
# Norm: Example

- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- HODGKIN'S DISEASE
- Hodgkin's disease
- Hodgkins Disease
- Hodgkin's disease NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Hodgkins disease
- hodgkin's disease
- Disease;Hodgkins
- Disease, Hodgkin
- ...

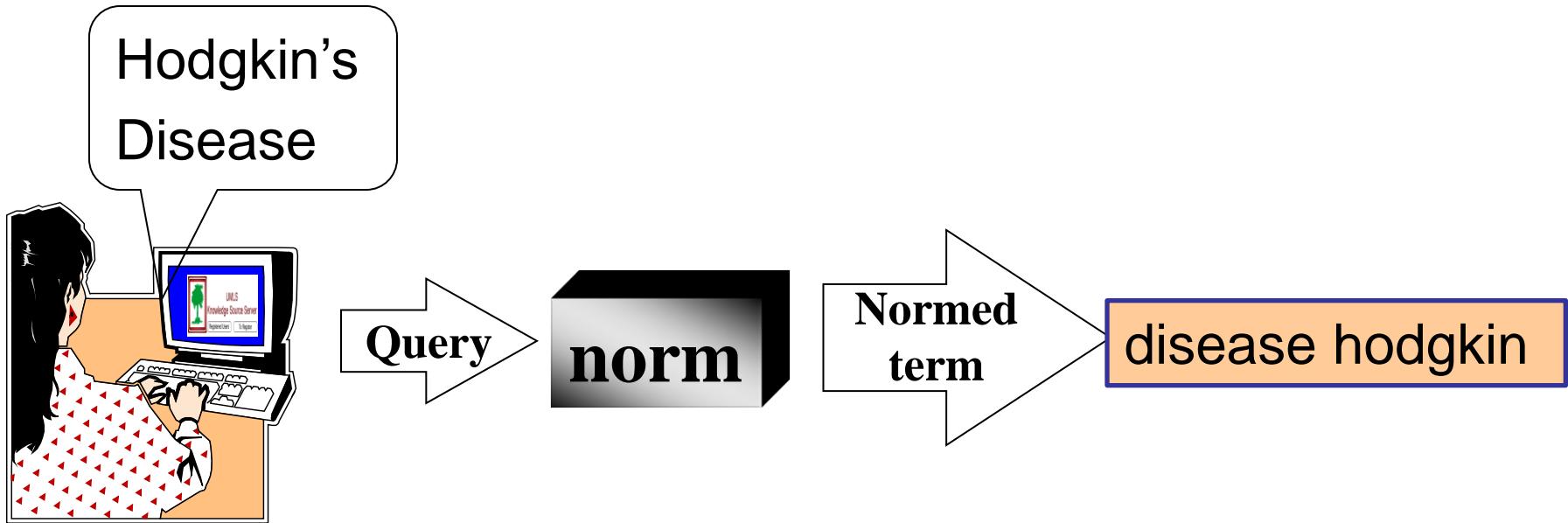


disease hodgkin

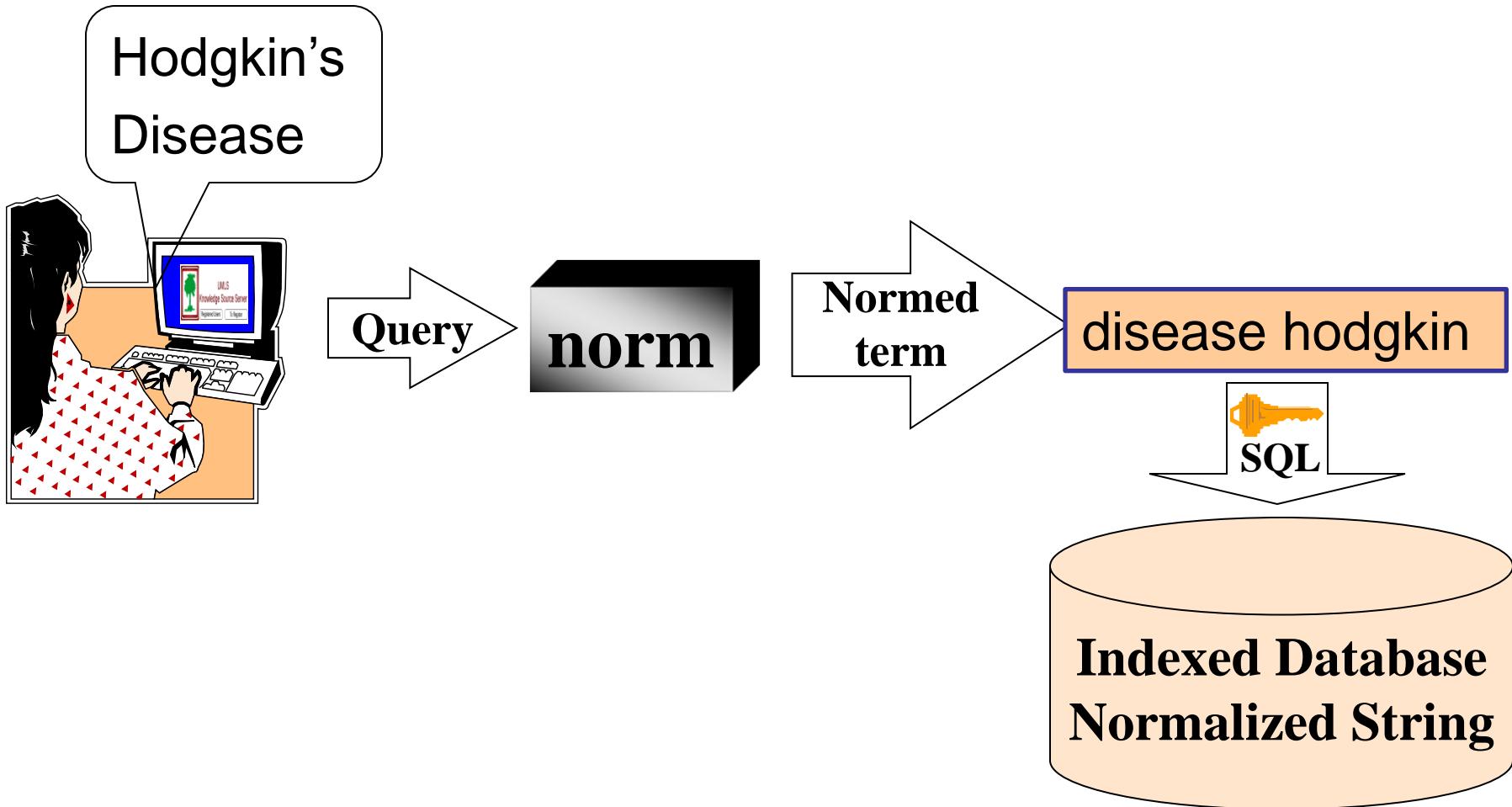
# Example - Norm



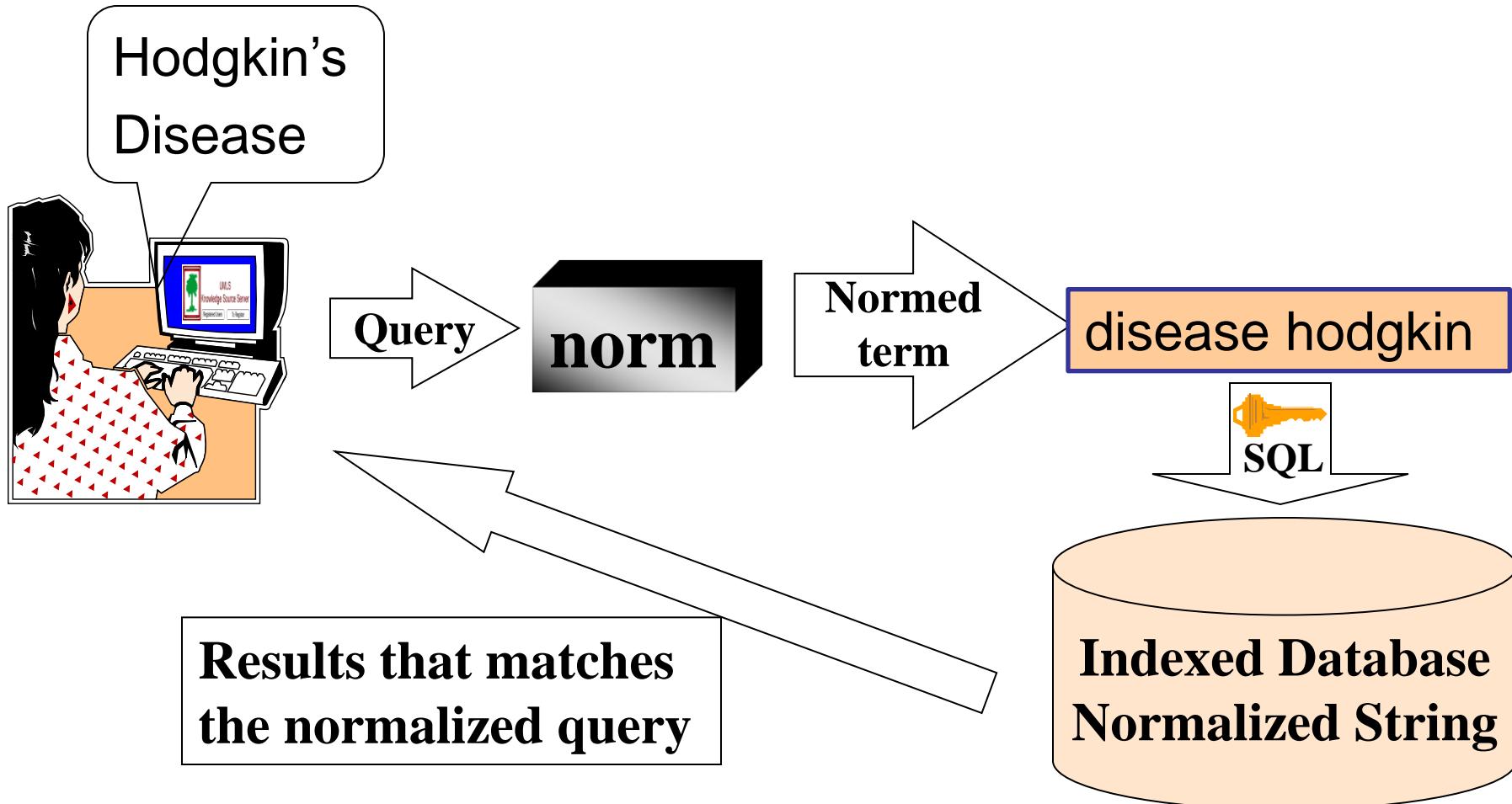
# Example - Norm



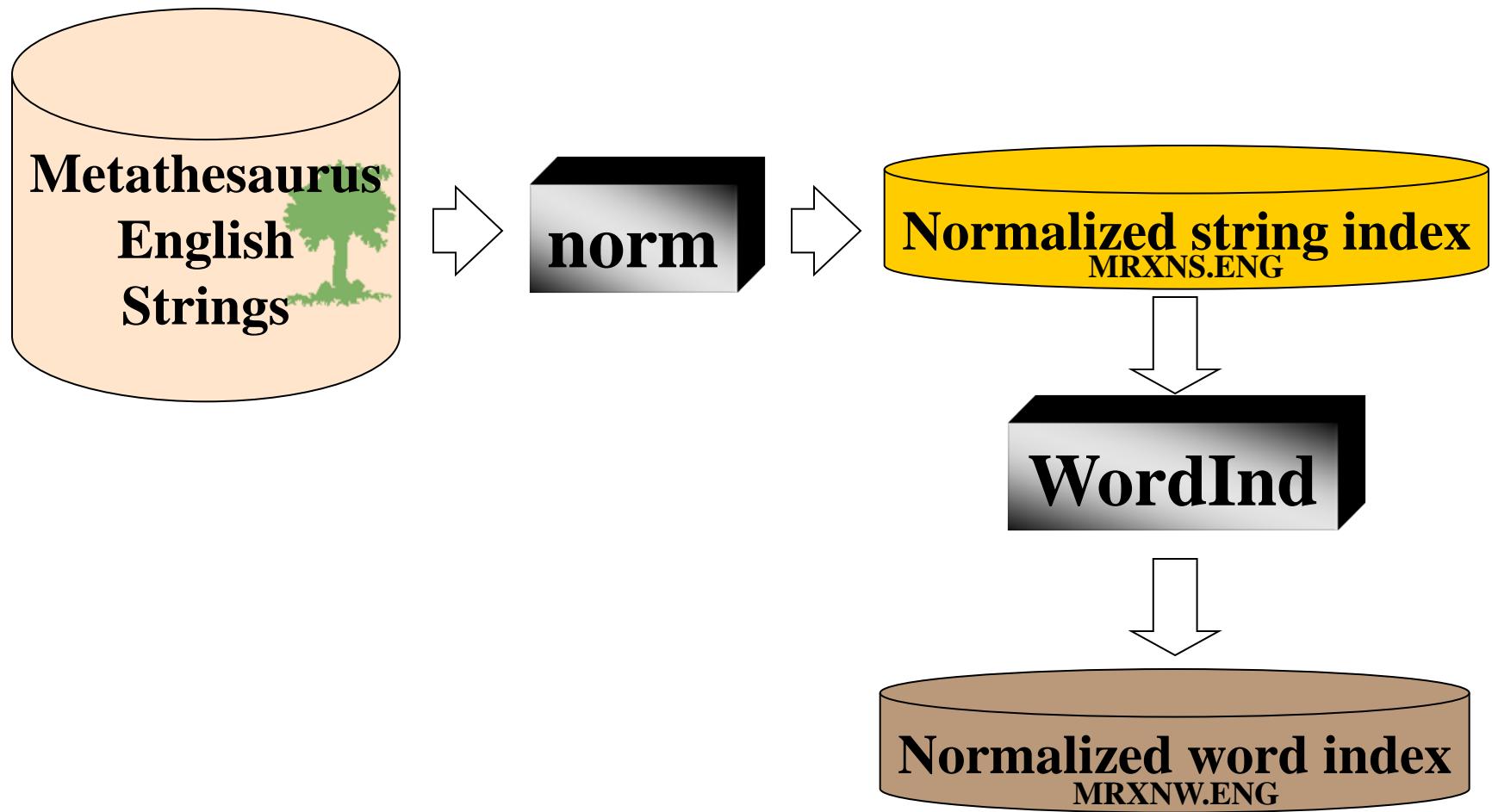
# Example - Norm



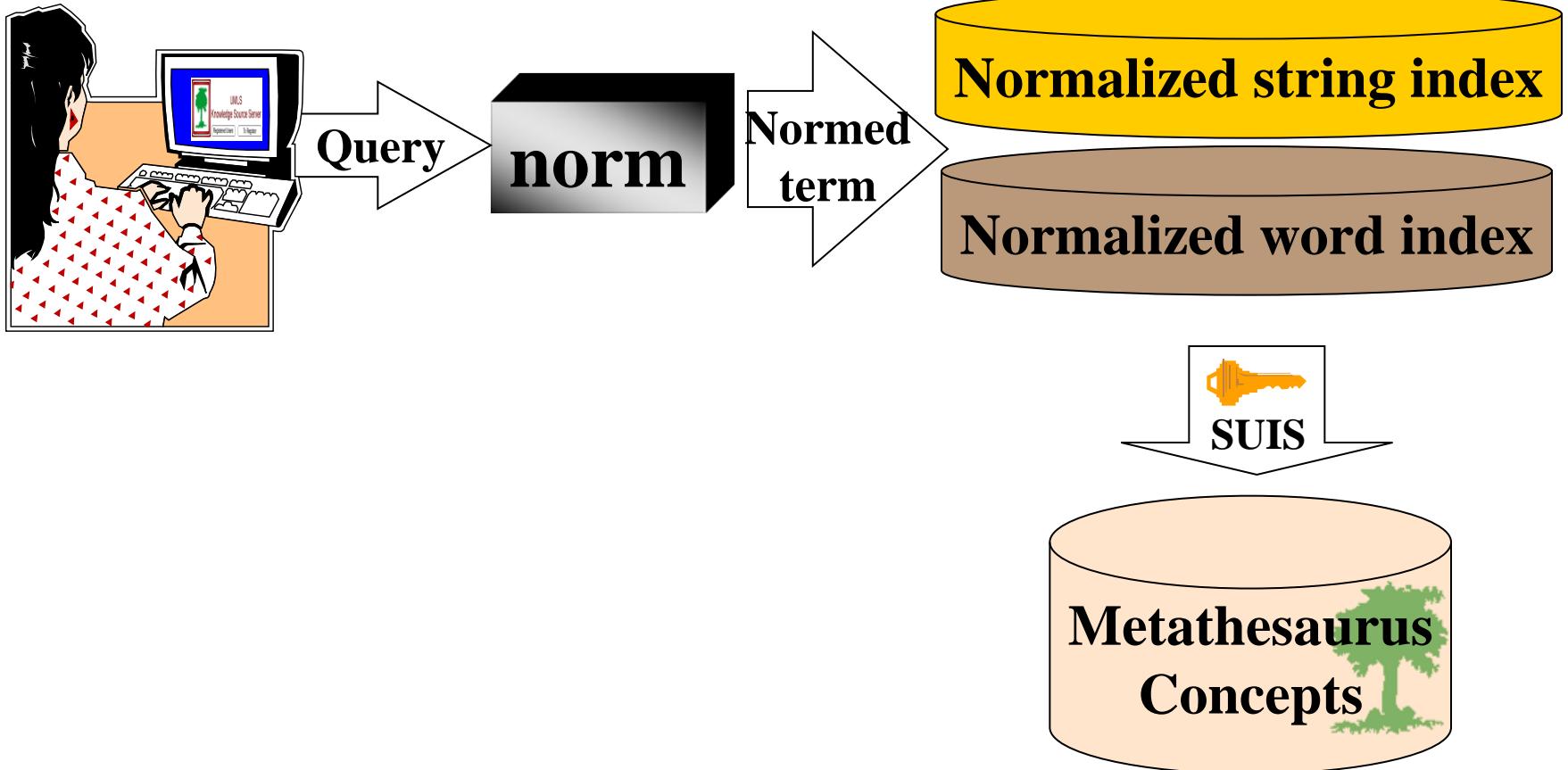
# Example - Norm



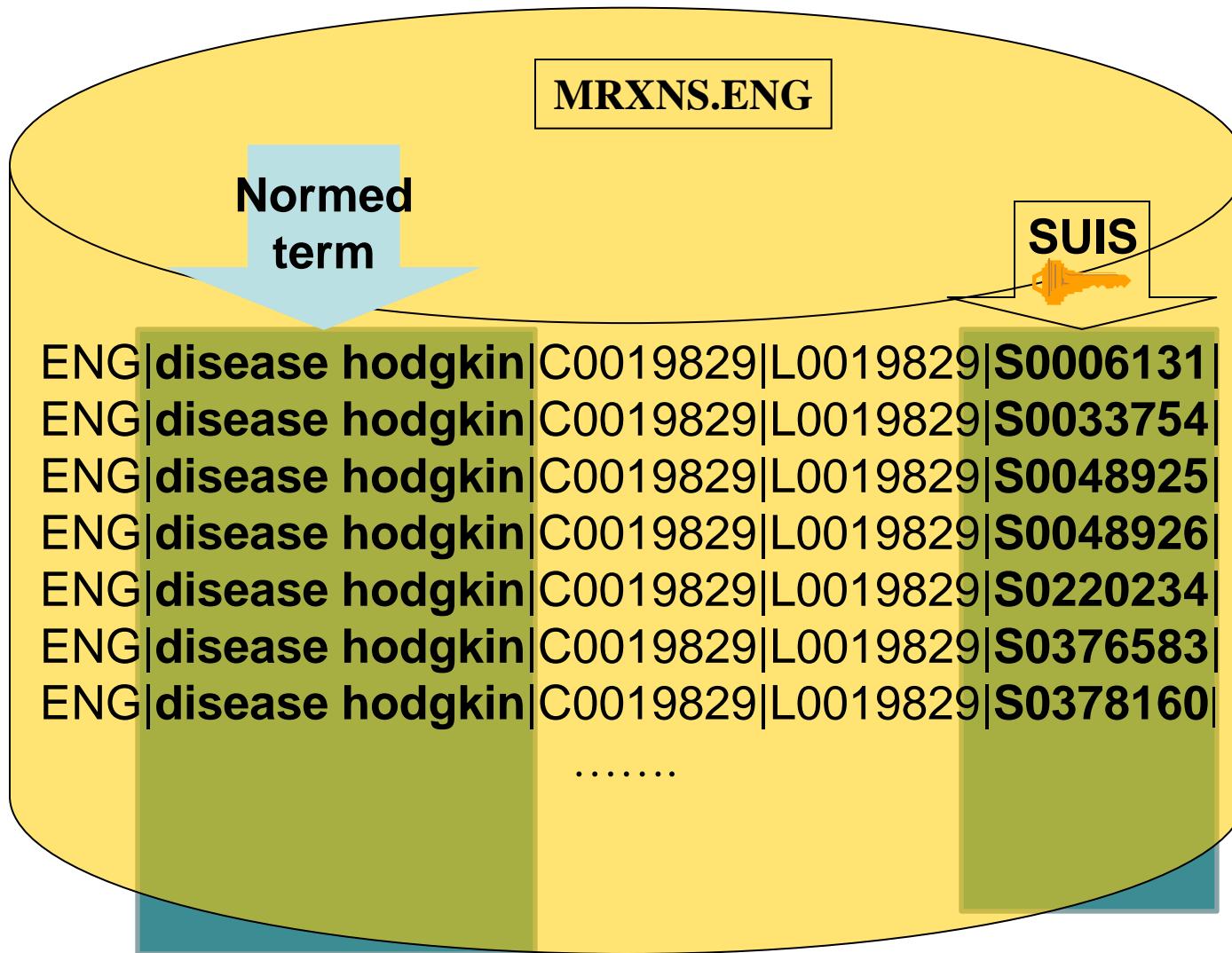
# Example 2 – UMLS Metathesaurus



# Example 2



# Example 2 – String Name



# Example 2 – String Name

MRCON

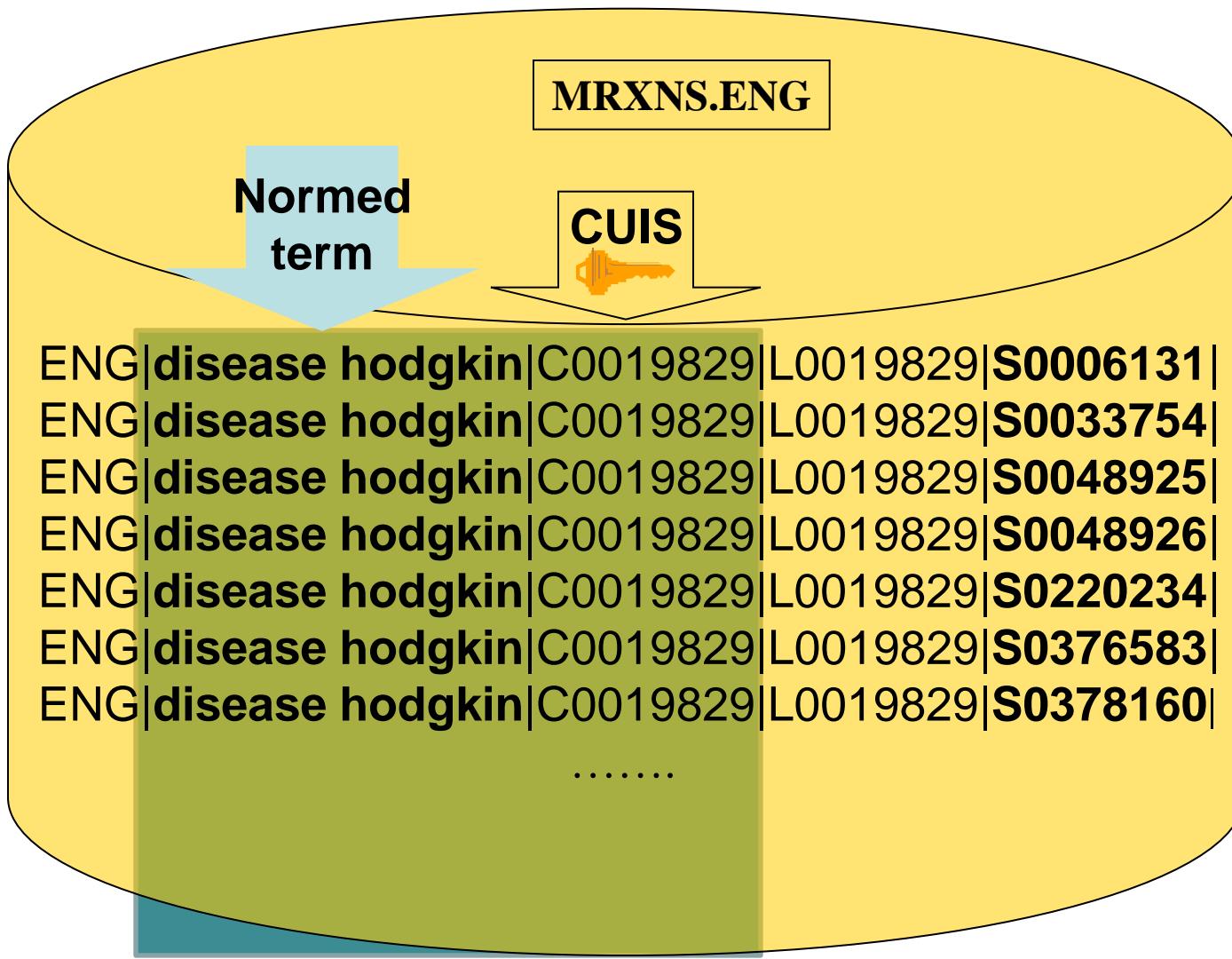


SUIS

C0019829|ENG|P|L0019829|**PF**|S0378161|Hodhkins Disease  
C0019829|ENG|P|L0019829|VC|S0006131|HODGKINS DISEASE  
C0019829|ENG|P|L0019829|VC|S0903124|Hodgkins disease  
C0019829|ENG|P|L0019829|VO|S0033574|Disease, Hodgkin  
C0019829|ENG|P|L0019829|VO|S0048925|Hodgkin Disease  
C0019829|ENG|P|L0019829|VO|S0048926|Hodgkin's Disease  
C0019829|ENG|P|L0019829|VO|S0220234|Disease, Hodgkin's

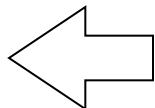
.....

# Example 2 – Concept Name



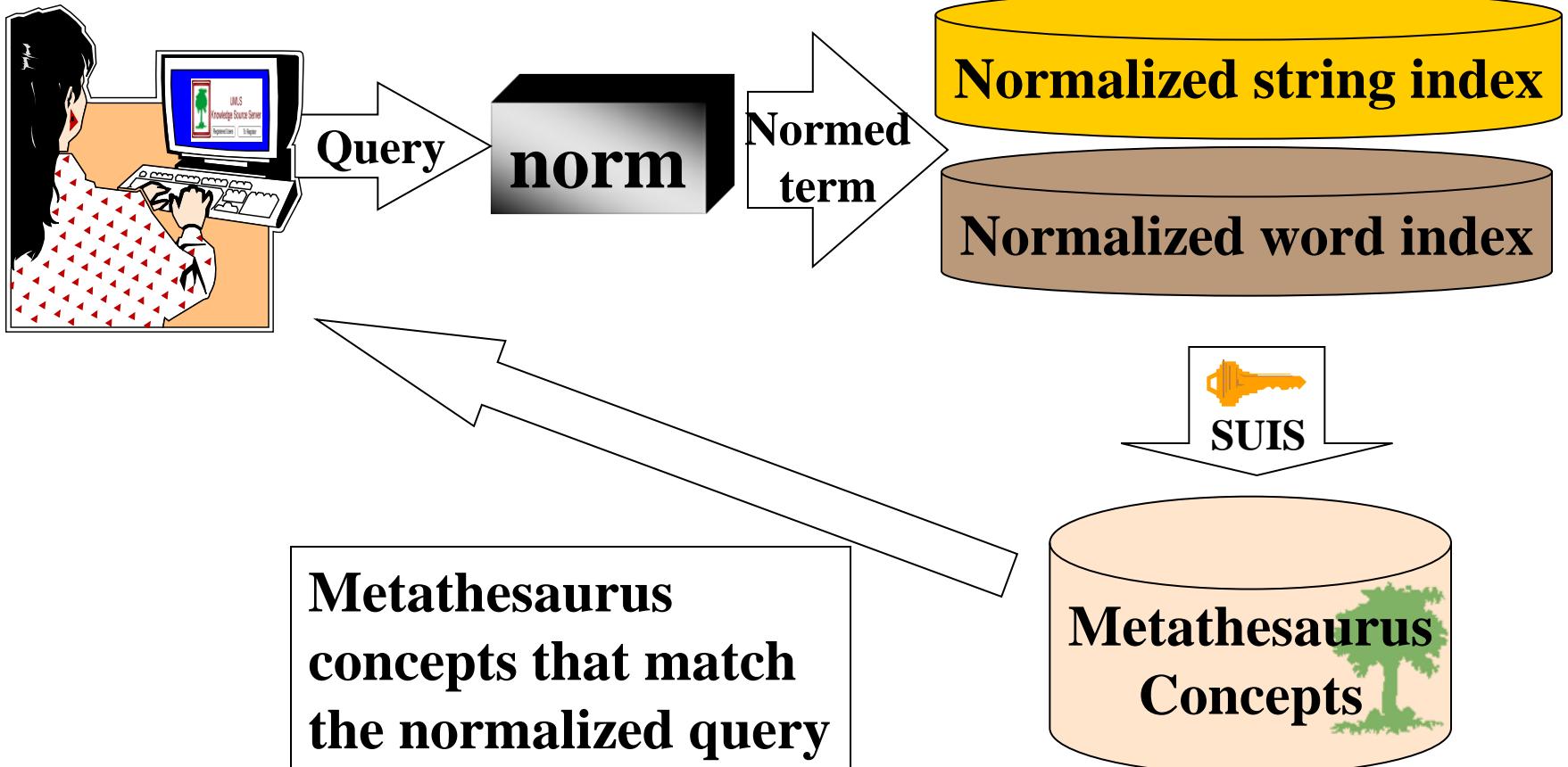
## Example 2

MRCON



C0019829 ENG P L0019829  <b>PF</b>  S0378161 Hodhkins Disease
C0019829 ENG P L0019829 VC S0006131 HODGKINS DISEASE
C0019829 ENG P L0019829 VC S0903124 Hodgkins disease
C0019829 ENG P L0019829 VO S0033574 Disease, Hodgkin
C0019829 ENG P L0019829 VO S0048925 Hodgkin Disease
C0019829 ENG P L0019829 VO S0048926 Hodgkin's Disease
C0019829 ENG P L0019829 VO S0220234 Disease, Hodgkin's
.....

# Example 2



# Text Categorization Tools



# TC Tools

- Based on Journal Descriptor Indexing (JDI) methodology (by Susanne Humphrey)
- Uses a small set of high level descriptors:
  - Journal Descriptors (JDs)
  - Semantic Types (STs)
- Used for categorizing text, indexing contents, retrieving records, and word sense disambiguation (WSD)

# Facts for TC Tools

- Release annually (since 2007)
- Free distributed with open source code
- 100% Java
- Run on different platforms
- One complete package
- Documents & supports
- Provides Java APIs, command line tools, GUI tools, and Web tools

# TC Tools

- Two types of categorization:
  - Journal Descriptor Indexing (JDI): categorizes text according to Journal Descriptors (JDs)
  - Semantic Type Indexing (STI): categorizes text according to Semantic Types (STs)
- St WSD tool (2009)

# Journal Descriptors (JDS)?

- Set of 122 MeSH descriptors representing high-level categories, mostly biomedical disciplines.
- Used for indexing journals *per se*
- Assigned by human indexer to the 4100 journals
- Source is from: List of Serials for Online Users file (Isi.xml)

# Journal Descriptors

- Examples of JD from Isi.xml
  - JID - 03132144
    - TA - Transplantation (the journal *Transplantation*)
    - JD - Transplantation
  - JID - 9802574
    - TA - Pediatr Transplant
      - (the journal *Pediatric Transplantation* )
    - JD - Pediatrics; Transplantation
  - JID - 0052631
    - TA - J Pediatr Surg (the *Journal of Pediatric Surgery*)
    - JD - Pediatrics; Surgery

# **JDI Methodology**

- Training set is about 3.4 million MEDLINE documents (3 years)
- JDI uses statistical associations between words in MEDLINE training set record TI/AB and the JD/s corresponding to the journal in the training set record
- But
  - JDs are not in a MEDLINE record
  - JDs are in the NLM serial record from Isi2007.xml

## JDI – Link to JDs

- Example of link between MEDLINE records and JDs
  - Training set MEDLINE record:  
PMID - 10919582  
TI - Combined liver and kidney transplantation in children.  
**JID - 0132144**  
SO - *Transplantation*. 2000 Jul 15;70(1):100-5.
  - *Transplantation* serial record:  
**JID - 0132144**  
JD - Transplantation

## **JDI – Link to JDs**

- Example of Training set MEDLINE record with “imported” JD Transplantation:

- **PMID - 10919582**

- TI - Combined liver and kidney transplantation in children.**

- SO - *Transplantation*. 2000 Jul 15;70(1):100-5.**

- JD - *Transplantation***

# JDI - JD Score (Word)

- JDI of the word “transplantation”

1|0.275691|Transplantation

2|0.070315|Hematology

3|0.044303|Nephrology

4|0.031517|Pulmonary Disease (Specialty)

5|0.029425|Gastroenterology

- Transplantation score

no. of docs in training set in which TI/AB

**word transplantation** co-occurs with **JD Transplantation**

=

no. of docs in training set in which the

**word transplantation** occurs in TI/AB

= 0.275691

## JKI - JD Score (Word)

- JDI of the word “kidney”

1|0.140088|**Nephrology**

2|0.080848|Transplantation

3|0.057162|Urology

4|0.032341|Toxicology

5|0.024398|Pharmacology

- Nephrology score

no. of docs in training set in which TI/AB  
**word kidney** co-occurs with **JD Nephrology**

$$= \frac{\text{no. of docs in training set in which the} \\ \text{word kidney occurs in TI/AB}}{\text{no. of docs in training set in which the} \\ \text{word kidney occurs in TI/AB}}$$

$$= 0.140088$$

## JKI - JD Score (Phrase)

- JDI of the phrase “kidney transplantation”

1|0.178269|**Transplantation**

2|0.092195|Nephrology

3|0.037875|Hematology

4|0.034381|Urology

5|0.017438|Gastroenterology

- Score for **Transplantation** is **average** of  
Transplantation score for **word kidney** and  
Transplantation score for **word transplantation**
- A JD score for a phrase is the average of that JD's  
score across the words in the phrase

# **STI - Semantic Types**

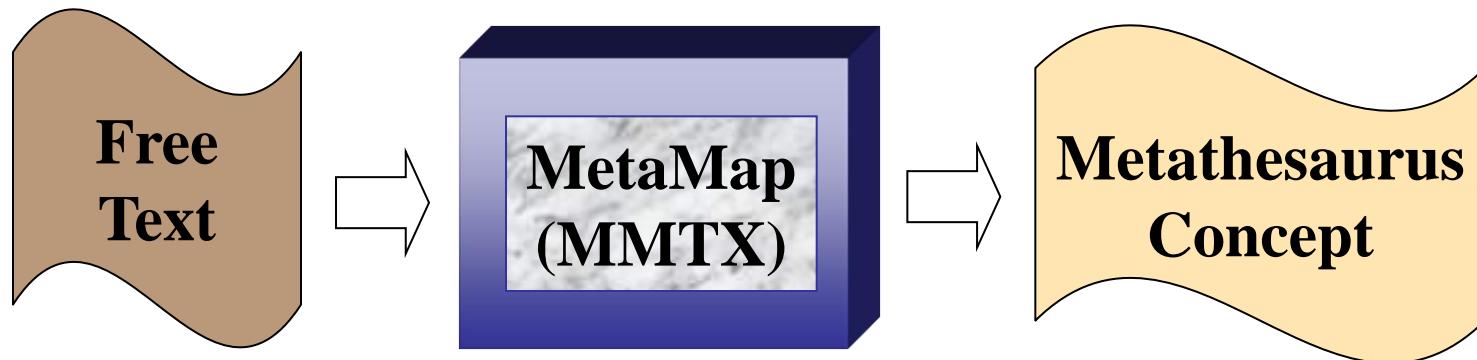
- What are Semantic Types (STs)?
- Set of 135 semantic types in the Semantic Network in NLM's Unified Medical Language System (UMLS).
- Concepts in the UMLS Metathesaurus are assigned one or more STs which semantically characterize those concepts
- For example, “aspirin” is assigned the STs Pharmacologic Substance (phsu) and Organic Chemical (orch).

## Semantic Type Indexing (STI)

- JDI has word-JD vectors representing JD indexing of each of the 304,000 words in the training set.
- STI also has word-ST vectors representing ST indexing of each training set word.
- Thus, STI of text can be performed exactly the same way as JDI of text. An ST score for a text is the average of that ST's score for words in the text. The scores for all the STs comprise the ST vector for the text.

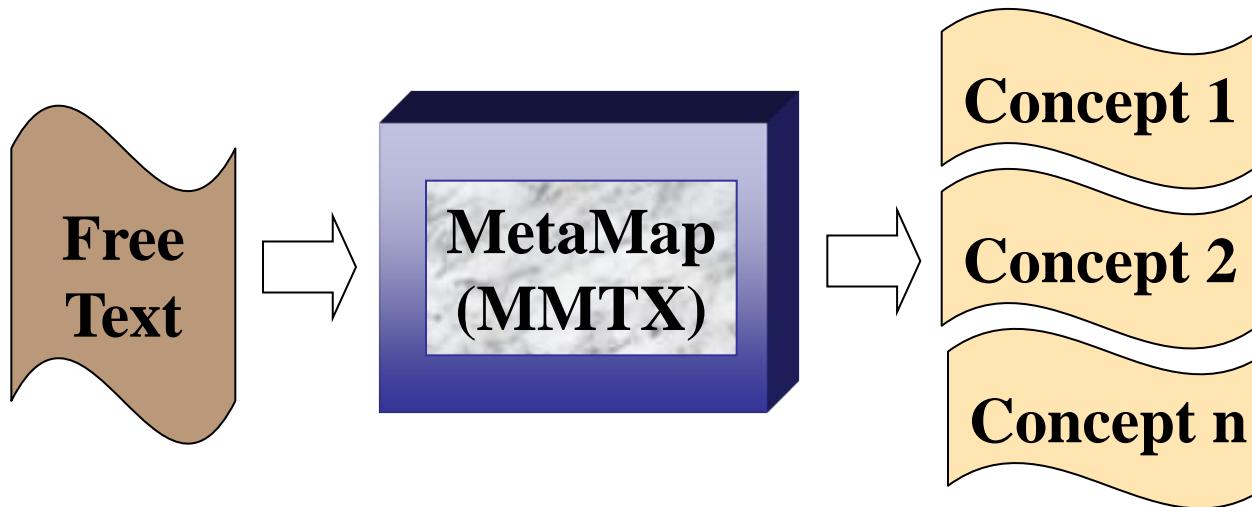
# TC – St WSD

- Words Senses disambiguation (WSD)



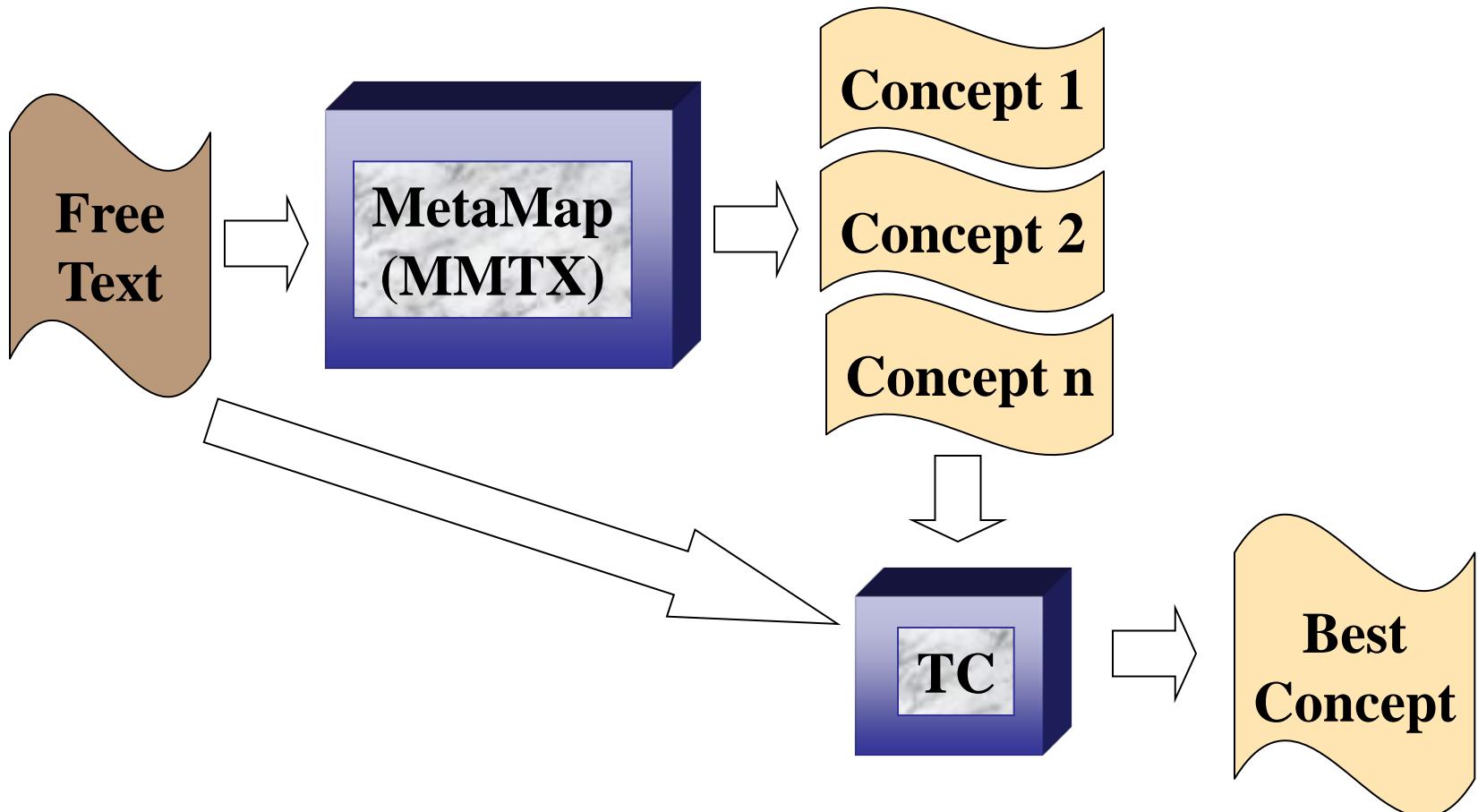
# TC – St WSD

- Words Senses disambiguation (WSD)



# TC – St WSD

- Words Senses disambiguation (WSD)



## Example – St WSD

- “transport” is ambiguous:
  - Biological Transport (ST is Cell Function, **celf**)
  - Patient Transport (ST is Health Care Activity, **hlca**)
- STI of text results in ranked list of STs.
  - If **celf** ranks higher than **hlca**, then meaning is Biological Transport.
  - If **hlca** ranks higher than **celf**, then meaning is Patient Transport.

# Example – St WSD

STI of PMID 9674486 in WSD collection

Input: Preliminary results of bedside inferior vena cava filter placement: safe and cost-effective. The use of inferior vena cava filters (IVCFs) is increasing in patients at high risk for venous thromboembolism; however, there is considerable controversy related to their cost. We inserted eight percutaneous IVCFs at the bedside. The hospital charges for bedside IVCF insertion were substantially lower compared with those for IVCF insertion performed in the Radiology Department or operating room. There was one death (unrelated to the procedure) and one asymptomatic caval occlusion believed to be caused by thrombus trapping. Bedside IVCF insertion is safe and cost-effective in selected patients. This practice averts the potential complications associated with **transporting** critically ill patients.

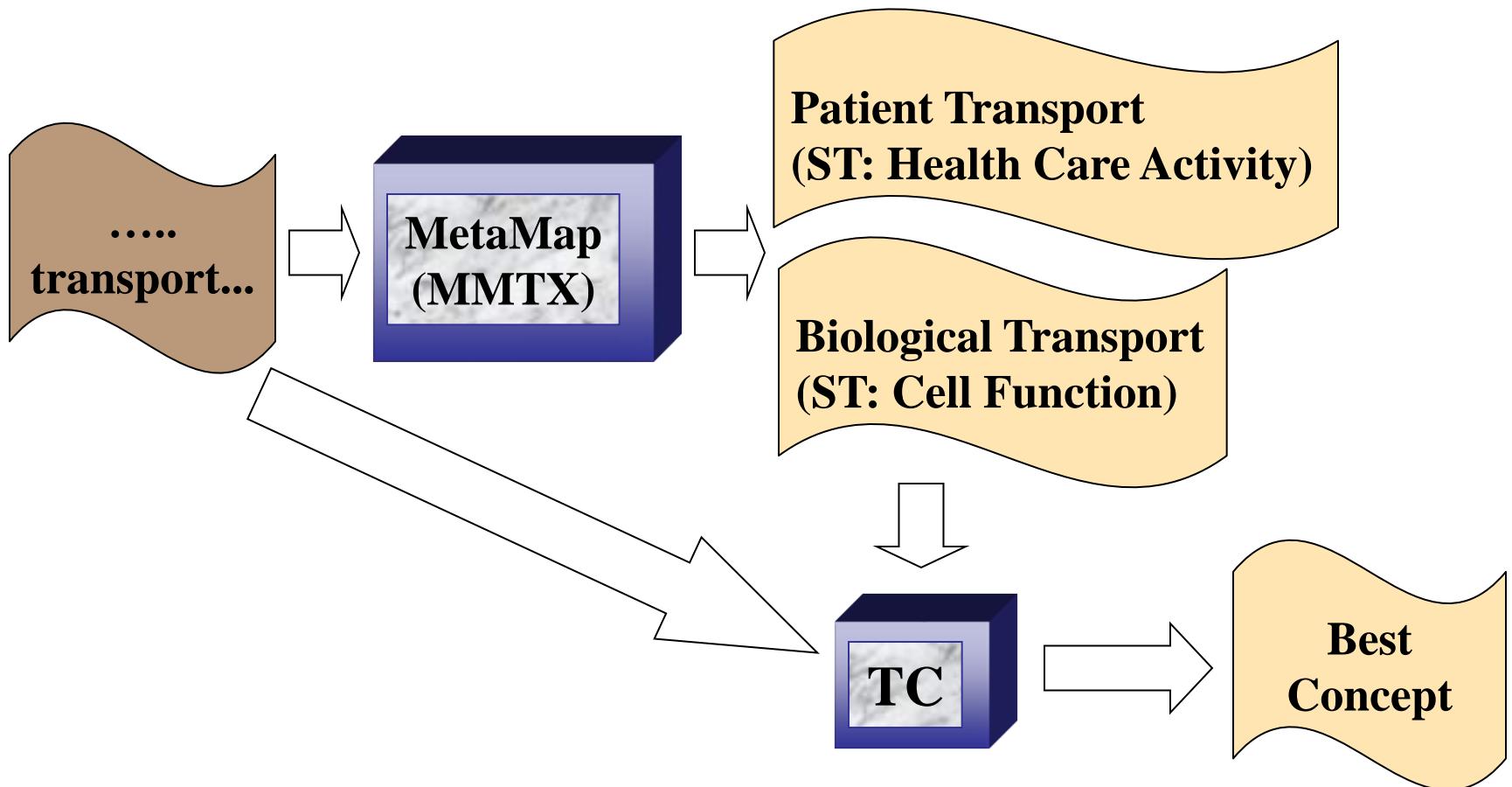
--- ST scores and rank based on document count for word ---

**27|0.4897|hIca|Health Care Activity <= Patient Transport**

46|0.4086|celf|Cell Function (~~Biological Transport~~)

# TC – St WSD

- Words Senses disambiguation (WSD)



## **TC – St WSD**

- Three methods for contexts of the ambiguity:
  - ambig-sentence - sentence with ambiguity
  - ambig-sentences - all sentences with ambiguity
  - doc - entire MEDLINE document
- Three score systems:
  - DC: document count
  - WC: word count
  - CS: combines score

## **TC – St WSD**

- Published research on STI as a tool for word sense disambiguation (WSD) in natural language processing (NLP) using UMLS Metathesaurus, disambiguating 45 ambiguous strings from NLM's WSD collection.
- Best unsupervised WSD methods
  - 2007: 75.39%
  - 2008: 75.00%
  - 2009: 77.37%
  - 2010: 77.36%
- First release in 2009.

# Questions



- Lexical Systems Group: <http://umlslex.nlm.nih.gov>
- The SPECIALIST NLP Tools: <http://specialist.nlm.nih.gov>